Deriving Association between Urban and Rural Students Programming Skills

[1]L.Arockiam, [2]S.Charles, [3]I.Carol, [4]P.Bastin Thiyagaraj, [5]S. Yosuva, [6]V. Arulkumar

[1]Associate professor, Department of Computer Science, St Joseph's College, Trichirappalli

[2]Lecturer, Department of Computer Science, St Joseph's College, Trichirappalli

[3, 4, 5, 6]Research Scholar, Department of Computer Science, St Joseph's College, Trichirappalli

Abstract - **Data mining is used to extract the interesting patterns from databases or repositories. Frequent Pattern Tree is a technique for discovering association between the variables and finds the frequent patterns in the Students' dataset. The student's data set contains the demographic details, practical mark and theory mark. The chore is to determine the programming skill similarity between rural and urban students. This paper focuses on two mining approaches to discover the knowledge from the students' dataset such as (a) Identifying Frequent Patterns (b) Finding similarity between urban and rural students programming skill**

Keywords - FP-Tree, K-means Clustering, Support and Confidence

I. INTRODUCTION

In data mining, Association rule mining is one of the approaches for finding the frequent patterns in the dataset. To find the useful knowledge, the data should be cleaned. If not, error can happen such as duplicated records and noisy data. The process will be done before mining process. The data mining technique is used to take out the useful information from the student data set. FP tree mining is used to determine all frequent item sets without the candidate's generation. The illiMine tool finds the frequent item set and constructs the tree using FP-Tree Technique. The frequent patterns are used as an input for the k-means clustering. It is used to categorize the frequent patterns based on the programming skill and place of living. This process is carried out using Weka Tool.

II. LITERATURE REVIEW

A. *Association Rule Mining*

Each attribute is treated as an item. There are three types of item values, which are used to for the attribute construction such as boolen and quantitative, categorical. The quality of a rule is defined by interesting problem based on the three categories such as 1. Generated rules can be self-evident. 2. Marginal events can dominate and 3. Interesting events can rarely occur. It is required to estimate how interesting the rules [2] are Subjective and Objective measures. Subjective measure is often based on earlier user experiences and belief. Objective measure is based on threshold values defined and controlled by the user.

Typical measures are 1.Support (utility) [1] is the percentage of transactions that demonstrate the rule. It defines usefulness of a rule which can be measured with a minimum support threshold. This parameter lets to measure how many events have such itemsets that match both sides of the implication in the association rule. 2. Confidence (certainty) [1] is certainty of a rule whichcan be measured with a threshold for confidence. This parameter lets to measure how often an event's itemset that matches the left side of the implication in the association rule and also matches the right side. It is the conditional probability that a given X present in a transaction implies the presence of Y. Confidence (X=>Y) equals Support (X, Y)/ Support (X). Association rule is derived from support and confidence. Rules are probabilistic in nature and are mined in two steps

1. Find all frequent item sets: Each item set occurs at least as frequently as a pre-determined min support count

2. Generating strong Association rules from frequent item sets. These rules must satisfy the min support as well as confidence.

B. FP-Tree Growth Algorithm

FP-growth algorithm is an efficient method of mining all frequent item sets without the candidate's generation. The algorithm mines the frequent item sets by using a divide-and-conquer strategy as follows: FP-growth first compresses the dataset representing frequent item set into a FP-tree, which retains the item set association information as well. The next step is to divide a compressed dataset into set of conditional dataset, each associated with one frequent item. Finally, it mines each such dataset separately. Particularly, the construction of FP-tree and the mining of FP-tree are the main steps in FP-growth algorithm. A frequent pattern tree is a tree structure consisting of an item-prefix-tree and a frequent item-header table. **Item-prefix-tree:** a) It consists of a root node labeled null b) Each non-root node consists of three fields such as Item name, Support count and Node link. **Frequent-item – header-table:** It consists of two fields such as Item name and Head of node link which points to the first node in the FP-tree [1, 2]

.C. *K-Means*

The K-means algorithm takes the input parameter k and partitions a set of n objects into k clusters so that the resulting intracluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity .

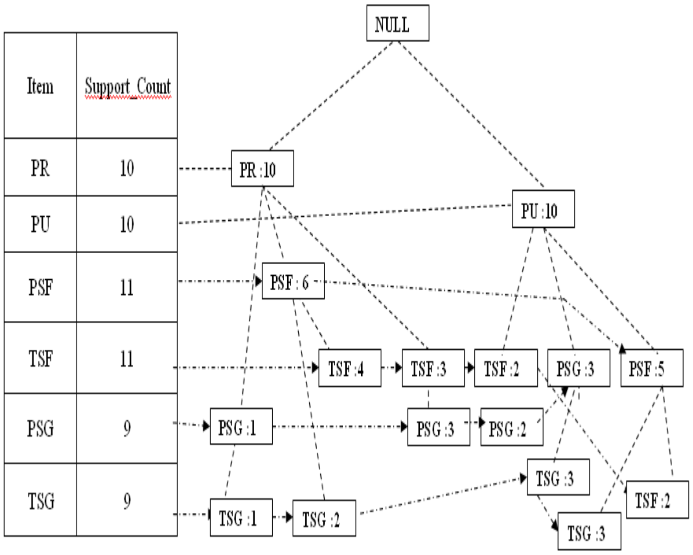
First, it randomly selects k of the objects, each of which initially represents a clusters mean or center. For each of the remaining objects, an object is assigned to the clusters to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

Where E is the sum of the square error for all objects in the data set; **p** is the point in space representing a given object; and **m$_i$** is the mean of cluster **c$_i$** (both **p** and **m$_i$** are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible [6].

III. METHODOLOGY

The students dataset contains the impute values such as theory mark (Assignment mark, Class test mark, Seminar mark , Viva mark, Mid semester mark and End semester mark) and practical mark(Mid semester mark and End semester mark). In this paper, the programming skill is evaluated based on theory and practical marks scored by the student in the programming subject. The programming skill is categorized as good or fair [9].



Item	Support Count
PR	10
PU	10
PSF	11
TSF	11
PSG	9
TSG	9

Fig. 1 FP-Tree design of the experiment

The illiMine Tool scans the dataset and determines a set of frequent 1-itemsets (L), which also include their support count. The set L is sorted in the order of descending support count. This ordering is important since each path of FP-tree will follow it. If the minimum support count is 4, then the set

L= {(PSG, 30), (PSF, 28), (TSG, 25), (TSF, 29)}.

The FP-Tree is generated based on association using the illumine [3, 4, 5] Tool .Fig 1 shows the relationship between the impute values in the dataset, which are classified as good or fair. The discovered frequent patterns are used as inputs for the k-means clustering [7,8]. This technique is used to form the clusters based on the performance and place of living.

IV. RESULTS AND DISCUSSION

The frequent patterns are identified using illiMine tool. It mines the 1000 samples and finds out the 333 frequent patterns from the post graduate and under graduate students' dataset. The frequent patterns are used to find the correlation between the programming skill of rural and urban students.

The fig 2 shows the programming skill categorization of rural and urban students. It reveals the association among the dataset based on the scores of programming subject. The frequent patterns are used as input variables, which is used to classify the dataset based on place of living and programming skill.

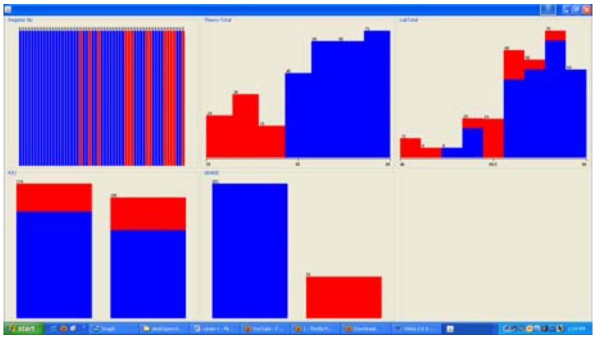


Fig. 2 Visualization of the Urban and Rural Students Categorization

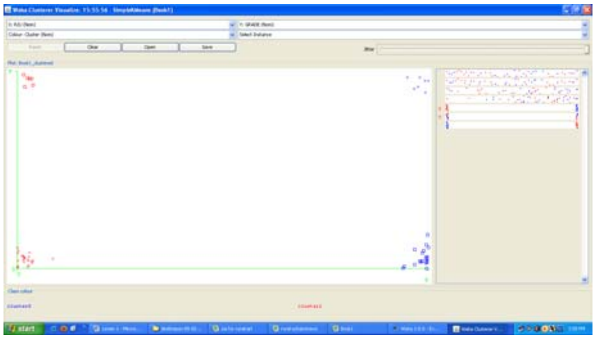


Fig. 3 Categorization of Clusters based on programming skill

Fig 3 shows the cluster representation of a grade and place of living in the students' dataset. It indicates that the most of the urban students are performing well in programming compared to rural students. Only few urban and rural students are scoring low marks in programming

```
=== Model and evaluation on training set
Number of iterations: 3
Within cluster sum of squared errors:
359.5807754918644
Clustered Instances
0      174 ( 53%)
1      156 ( 47%)
Class attribute: GRADE
Classes to Clusters:
   0   1  <-- assigned to cluster
 138 114 | G
  36  42   |  F
Cluster 0 <-- G
Cluster 1 <-- F
```

Fig. 4 Learning style preferences of Urban and rural students

Fig. 3, 4 shows two clusters such as cluster 0 and 1. Cluster 0 cloud holds the good programming skill student objects. Cluster 1 cloud holds the fair programming skill object. In cluster 0, 138 student objects are situated in urban cloud and 36 student objects are situated in rural cloud. In cluster 1, 114 student objects are situated in urban cloud and 42 student objects are situated in rural cloud. . Each cluster reveals the identity based on the programming skill.

The findings indicate that the more number of urban students are good in programming skills compared to rural students and more number of rural students are fair in programming skill.

V. CONCLUSION

In this paper, **FP Tree and K-means clustering technique** is used for finding the similarity between urban and rural students programming skills. FP Tree mining is applied to sieve the patterns from the dataset. K-means clustering is used to determine the programming skills of the students. This study clearly indicates that the rural and the urban students differ in their programming skills. The huge proportions of urban students are good in programming skill compared to rural students. It divulges that academicians provide extra training to urban students in the programming subject.

REFERENCES

[1] Aiman Moyaid SaidA, Dr. P D D. DominicB, Dr. Azween B AbdullahC **"**A Comparative Study of FP-growth Variations", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.5, May 2009.

[2] J. Han, J. Pei, and Y. Yin, " Mining Frequent Patterns without Candidate Generation", Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'00), 2000.

[3] O. Zaiane, M. El-Hajj, and P. Lu. Fast Parallel Association Rules Mining without Candidacy Generation. In *IEEE 2001 International Conference on Data Mining (ICDM'2001)*, 2001.

[4] Liu, L., Li, E., Zhang, Y., and Tang, Z. 2007. Optimization of frequent itemset mining on multiple-core processor. In Proceedings of the 33rd international Conference on Very Large Data Bases. 2007.

[5] C. Borgelt. An Implementation of the FP-growth Algorithm. Workshop Open Source Data Mining Software. 2005.

[6] Antonio R. Anaya, Jesús G. Boticario "A Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks", Educational Data Mining, 2009.

[7] Anaya, A.R., Boticario, J.G. Clustering Learners according to their Collaboration. 13th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2009.

[8] .Bratitis, T., Dimitracopoulou, A., Martínez-Monés, A., Marcos-García, J.A., Dimitriadis, Y. Supporting members of a learning community using interaction analysis tools: the example of the Kaleidoscope NoE scientific network Proceedings of the IEEE International Conference on Advanced Learning Technologies, ICALT 2008, 809-813, Santander, Spain, July 2008.

[9] Norazlina Khamis1 , Sufian Idris, "Investigating current object oriented programming assessment method in Malaysia's universities" , Proceedings of the International Conference onElectrical Engineering and Informatics Institute of Teknologi Bandung, Indonesia June 17-19, 2007.

[10] R.Agrawal, T.Imielinski and A.Swami, "Mining Association Rules between Sets of Items in Large DataBase", Proceeding of the ACM SIGMOD Conference, Washington, DC, 1993.

[11] Jiawei Han and Micheline Kamber, "Data Mining: concepts and techniques", Morgan Kaufmann Publishers, San Francisco, 2006.

[12] Romero C. and Ventura S, "Educational data mining: A Survey from 1995 to 2005".Expert Systems with Applications 33, pp.135-146, 2007.

[13] Sheikh L, Tanveer B. and Hamdani S, "Interesting Measures for Mining Association Rules", IEEE-INMIC Conference, December 2004.

[14] Dunn.R and Dunn.K Teaching Students through their learning styles: A practical approach. Reston, VA: Reston Publishing Company Inc.!978

Dr. L. Arockiam is working as Associate Professor in the Department of Computer Science, St.Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 21 years of experience in teaching and 13 years of experience in research. He has published 47 research articles in the International / National Conferences and Journals. He has also presented 2 research articles in the Software Measurement European Forum in Rome. He has chaired many technical sessions and delivered invited talks in National and International Conferences. He has authored a book on “Success through Soft Skills”. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining and Mobile Networks.

S.Charles is working as Lecturer in the Department of Computer Science, St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 13 years of experience in teaching. He has published many research articles in the National / International conferences and journals. He has acted as a chair person for many national and international conferences. He is currently pursuing doctor of philosophy programme and his current area of research is Data mining.

V.Arulkumar is pursuing doctor of philosophy in Department of Computer Science, St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He received his M.Phil degree from St.Joseph’s College, Tiruchirappalli. He received his M.Sc (Information Technology) from K.S.R College of Technology, Tiruchengode, Many Research article are published in the National / International conferences and journals. His current area of research is Data mining.

I. Carol is pursuing Master of philosophy in Department of Computer Science, St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He received his MCA degree from St.Joseph’s College, Tiruchirappalli. His area of interest is Data mining.

P. Bastin Thiyagaraj is pursuing Master of philosophy in Department of Computer Science, St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He received his MCA degree from St.Joseph’s College, Tiruchirappalli.

S. Yosuva is pursuing Master of philosophy in Department of Computer Science, St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He received his M.Sc(Computer Science & Information Technology) degree from G.T.N Arts and Science College, Dindigul.